Mapping large-area impervious surface and forest canopy density using Landsat 7 ETM+ and high resolution imagery

Limin Yang and Chengquan Huang
Raytheon ITSS
USGS EROS Data Center
Sioux Falls, SD

2002 High Spatial Resolution Commercial Imagery Workshop
March 25-27, 2002
Reston, Virginia, USA

Acknowledgements:

- National Land Cover Mapping Strategy Team (EDC)
- NASA SDP Program
- Earthsat Corporation

Background

- Multi-resolution Land Characteristics Consortium (MRLC) was initialed in early 1990s to address the need for consistently developed national and regional land cover data

Multi-Resolution Land Characteristics Consortium

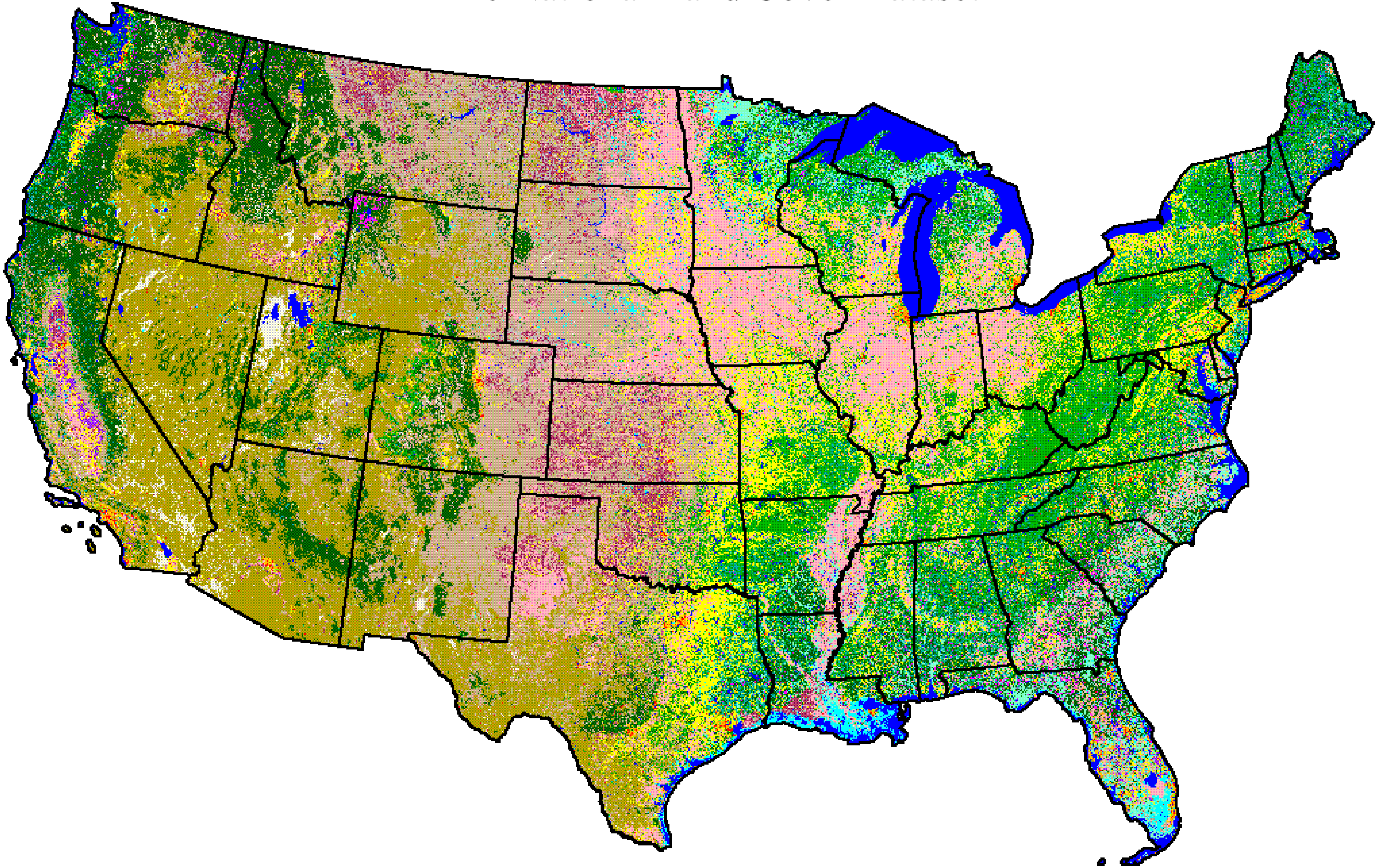
- Acquisition
  - Processing
    - EPA
      - Land Cover
    - EDC
      - Land Cover
    - NAWQA
      - Water Quality
    - GAP
      - Vegetation Cover
    - USFS
      - Inventory Fire Fuels
    - NOAA
      - Coastal Change

Background

- Through MRLC consortium, a 1992-vintage National Land Cover Dataset (NLCD) was developed for the conterminous United States

USGS
science for a changing world

The National Land Cover Dataset



U.S. Department of the Interior   U.S. Geological Survey

Sept 2000

USGS

Land Cover Characterization Program

National Land Cover Classification System

- Analogous to Anderson 1-2
- Merging of other current systems
- 21 “Anderson” Classes

National Land Cover Dataset Classes

- Open Water
- Perennial Ice/Snow
- Low Intensity Residential
- High Intensity Residential
- Commercial/Industrial/Transportation
- Bare Rock/Sand/Clay
- Quarries/Strip Mines/Gravel Pits
- Transitional
- Deciduous Forest
- Evergreen Forest
- Mixed Forest
- Shrubland
- Grasslands/Herbaceous
- Orchards/Vineyards
- Pasture/Hay
- Row Crops
- Small Grains
- Fallow
- Urban/Recreational Grasses
- Woody Wetlands
- Emergent Herbaceous Wetlands

Proposed NLCD 2000

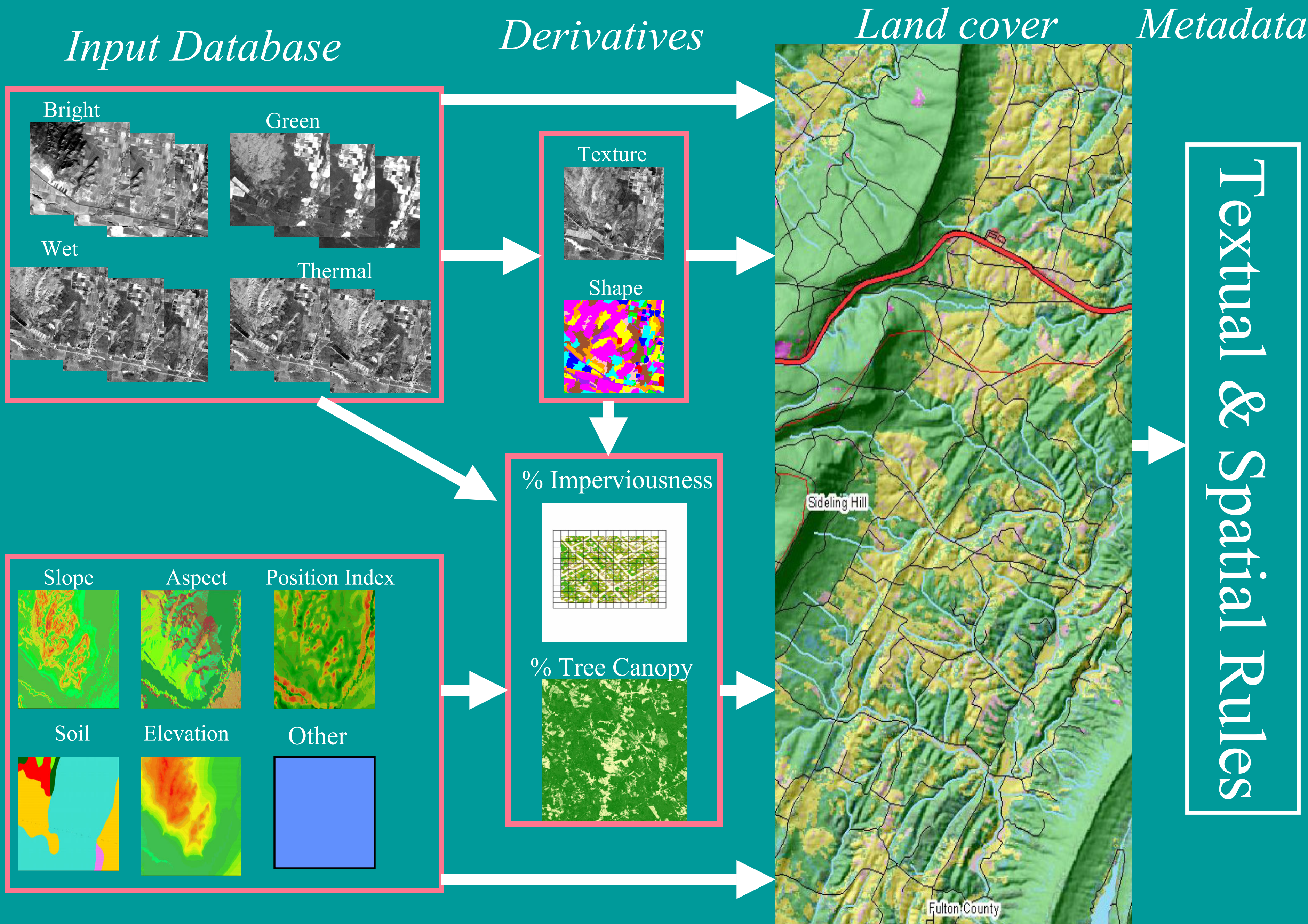
Mapping using Landsat 7 imagery

- 50 States and Puerto Rico
- Begin in FY 2000, completion TBD

Database approach

- Three Landsat 7 ETM+ scenes for each path/row (radiometric, geometric and terrain corrected and referenced to Albers Equal area projection)
- Land cover type of 30 meter resolution (categorical)
- Sub-pixel imperviousness estimate (continuous)
- Sub-pixel tree canopy density estimate (continuous)
- Shape/texture
- DEM and derivatives
- Other geospatial ancillary data

NLCD 2000 Database



Research objectives:

- to develop a repeatable, reasonably accurate, and cost-effective method to map sub-pixel percent impervious surface and tree canopy density at 30-meter resolution for the United States

Impervious surface:

• any impenetrable surface that prevents infiltration of water into it, such as:

rooftops, roads and parking lots, sidewalks

Source: Space Imaging

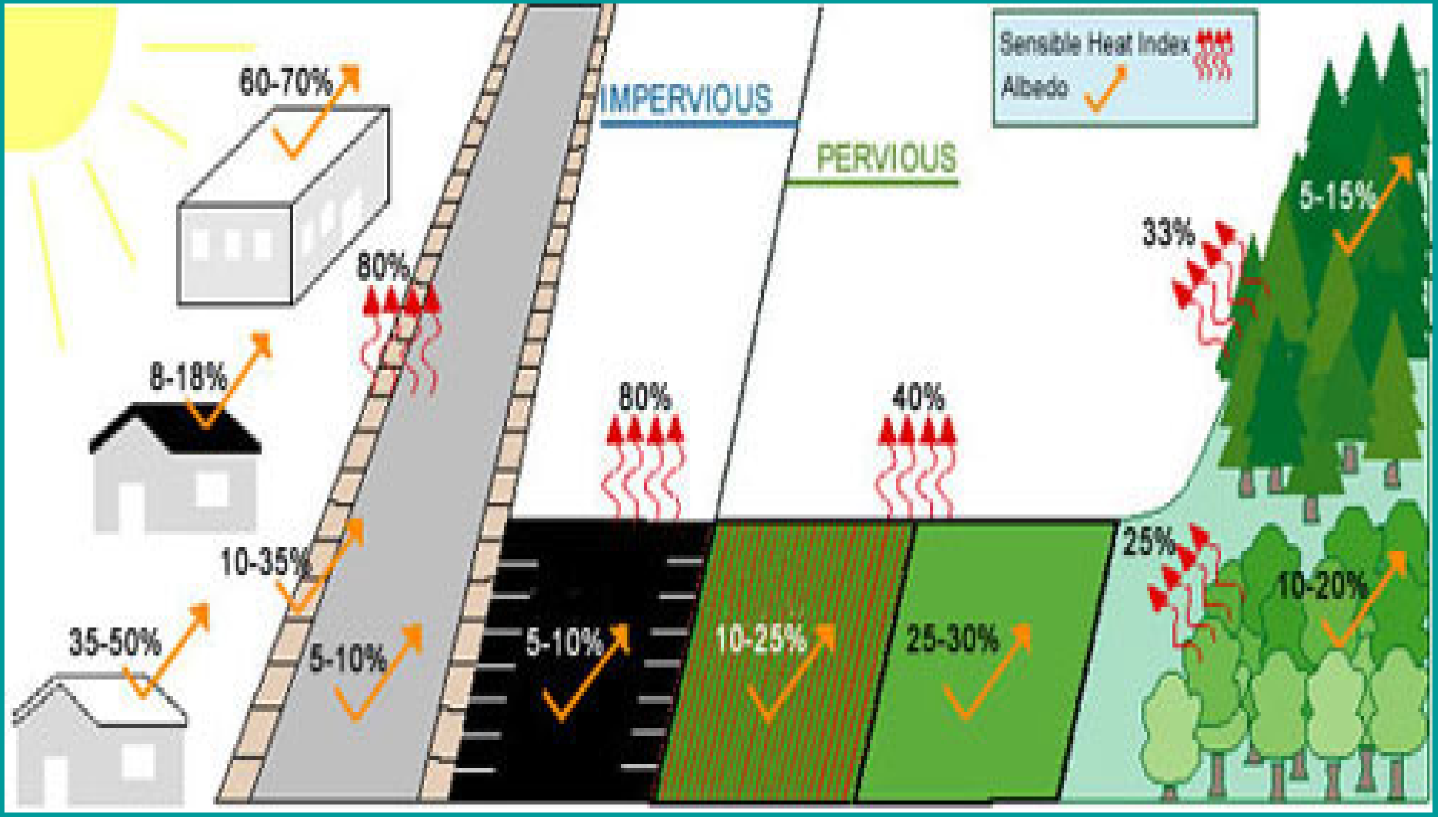
Why model impervious surfaces?

- One of the key indicators to characterize urban suburban land cover and land use and environmental conditions

- Wide range of potential applications in:

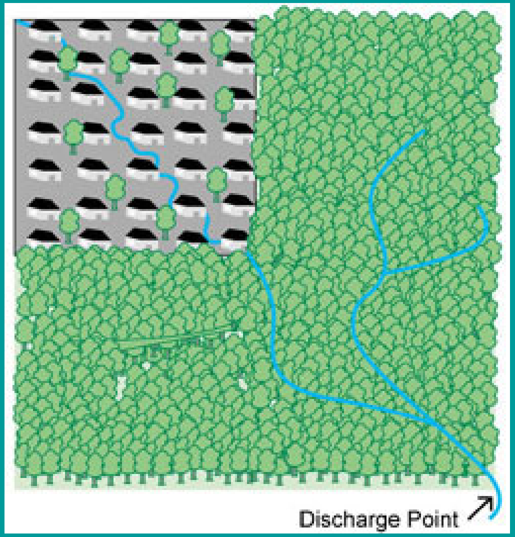
  land cover land use characterization
  urban hydrology
  urban climate
  urban planning
  urban pollution
  habitats and aesthetics

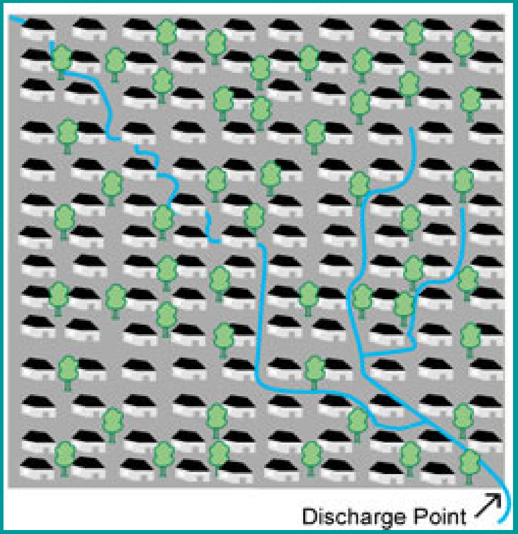
Impact on micro-climate

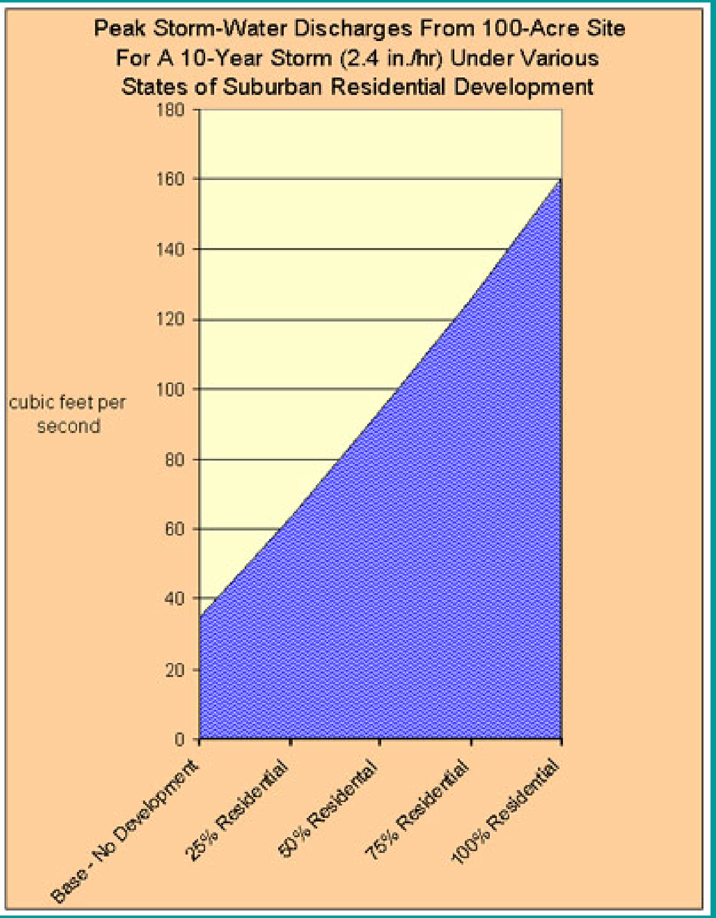


Source: The Chesapeake Bay from Space project
the Towson University Center for
Geographic Information Sciences (CGIS)

Impact on water quantity and quality







Source: The Chesapeake Bay from Space project
the Towson University Center for
Geographic Information Sciences (CGIS)

Forest type and canopy density mapping

- Classification – discrete or continuous?
- IGBP classification scheme:
    - Forest classes: canopy cover > 60%
    - Woody savannah: canopy cover between 30 and 60%
    - Savannah: canopy cover between 10 and 30%
    - Non-forest classes: canopy cover < 10%
  - Land cover often varies continuously over space
  - Different schemes often use different threshold values

Methods for estimating subpixel land cover

- **Physically-based models** (e.g. Li and Strahler,1992)
    - May be too complex to be inverted for large-area application
- **Spectral mixture models (**e.g. Martin, 2000; Flanagan and Civco, 2001)
    - End-members – green vegetation, non-photosynthetic vegetation, soil etc.
- **Statistical models**:
  - Linear regression (e.g. Ridd, 1995)
    - can not approximate non-linear relationships
    - Neural net (e.g. Civco et al.1997)

Modeling method (this study):

- A Regression Tree algorithm:

    a machine-learning algorithm

    recursively partitions data samples into subsets

    develops a linear model for each subset

CUBIST (Rulequest Research Inc.)

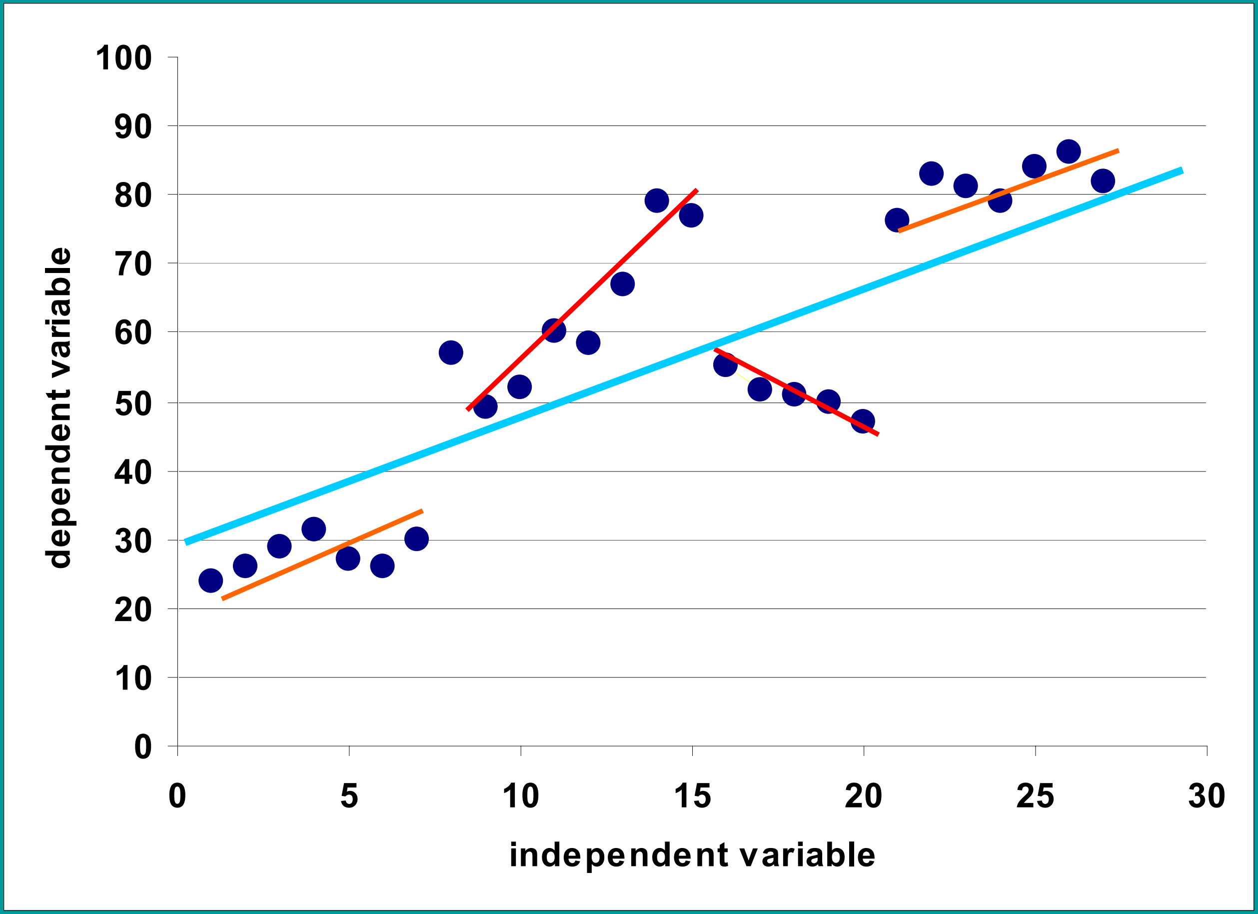
Model method (cont.)

Advantages:

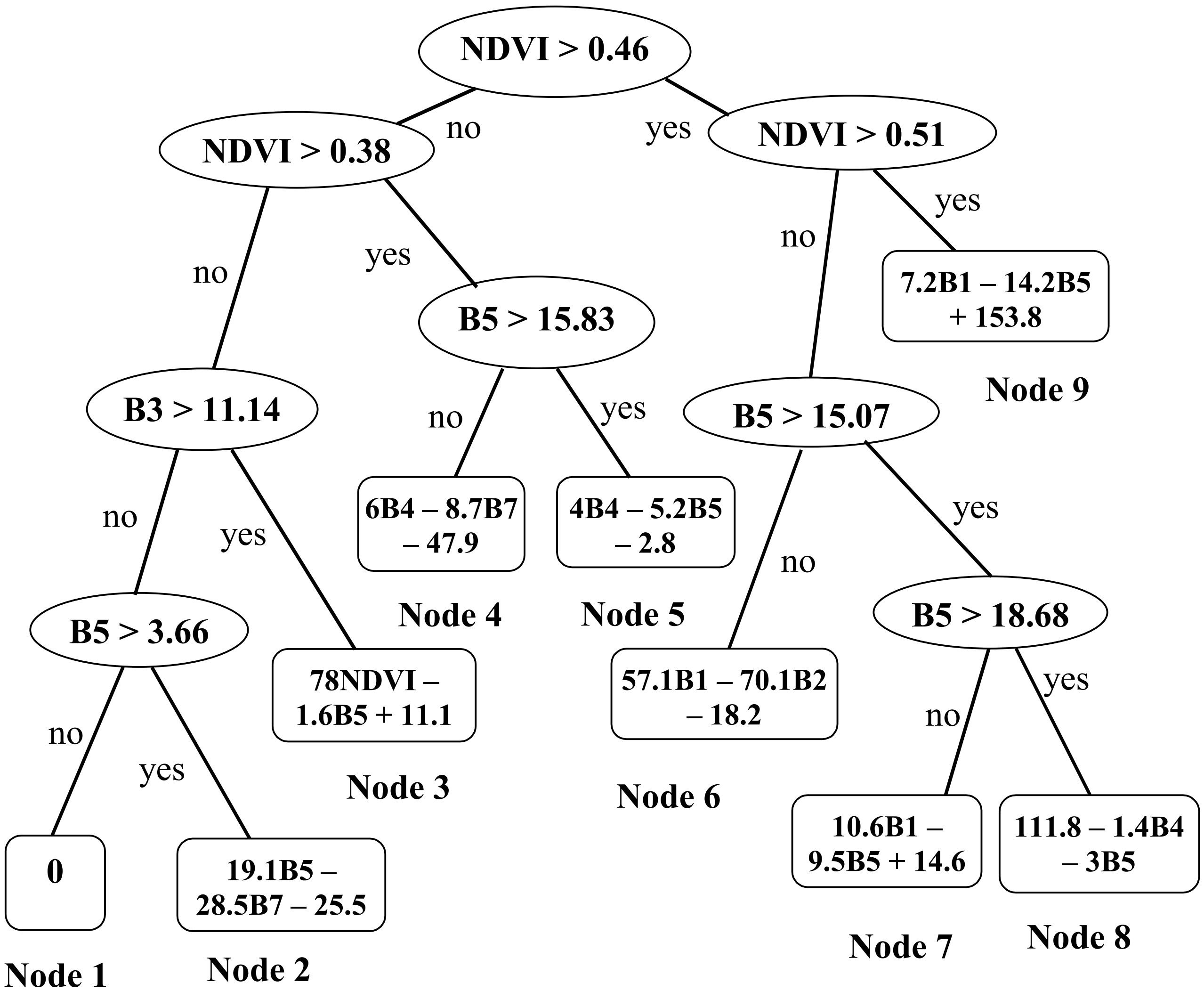
- categorical and numerical data input
- approximate complex nonlinear relationships
- rules generated are interpretable
- repeatability

Limitations:

- highly dependent on training data quality
- over-fitting


100
90
80
70
60
50
40
30
20
10
0
dependent variable
0
5
10
15
20
25
30
independent variable

Model development – an example regression tree model



Regression tree:

- Recursively partitions data samples into subsets
- Develops a linear model for each subset
- Minimizes the overall residual sum square of error
- Can approximate complex nonlinear relationships

Model

Rule 1: [12 cases, mean 20.4366207, range 0.288889 to 49.55042, est err 10.9970322]

if

tmband4 > 61

NDVI > 0.0619469

then

percent impervious = 88.3936 - 1.016 tmband4 + 0.44 tmband3 - 31.7 NDVI

Rule 2: …,…

Regression Tree method

The measure of best split at the node is based on the impurity of an example set. The expression for measuring impurity can be defined as (Karalic, 1992):

$$I(E) = \frac{1}{W(E)} \sum_{e_i \in E} (y_i - g(\vec{x}_i))^2$$

Function $g(\vec{x}_i)$ represents the regression plane through the example set.
Expected impurity of a split is estimated as

$$I_{\exp} = p_l I_l + p_r I_r$$

Where p_l, p_r denote probabilities of transitions into the left and the right son of the node, and I_l, I_r are corresponding impurities.

The quality of the constructed regression tree can be measured by the **mean absolute error R of a tree T**, expressed by

$$R(T) = \frac{1}{N}\sum_{i=1}^{n} | y_i - g(x_i) |$$

where *N* is the number of examples used for test.

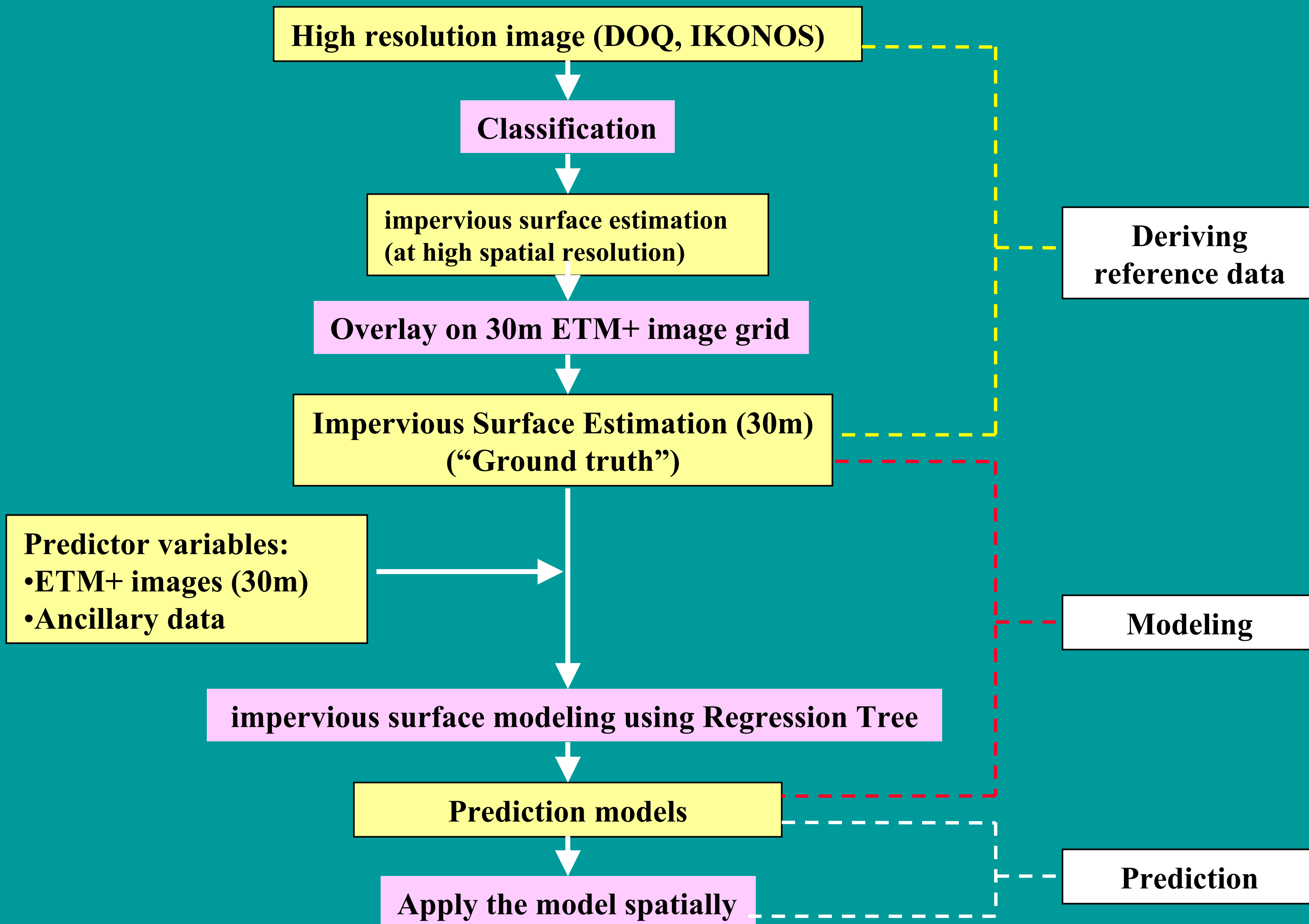
To compare the quality of several trees, the **relative mean absolute error** is often used and is defined as:

$$RE(T) = \frac{R(T)}{R(\mu)}$$

where $R(\mu)$ is the mean absolute error of the predictor which always predicts the mean value of the training example data set. It is used here to standardize the $R(T)$.

Procedures for imperviousness/canopy mapping:

- training data development
- modeling using regression tree algorithm
- spatial mapping/predicting


High resolution image (DOQ, IKONOS)
Classification
impervious surface estimation
(at high spatial resolution)
Overlay on 30m ETM+ image grid
Impervious Surface Estimation (30m)
("Ground truth")
Predictor variables:
•ETM+ images (30m)
•Ancillary data
impervious surface modeling using Regression Tree
Prediction models
Apply the model spatially
Deriving reference data
Modeling
Prediction

Study Areas:

within the United States representing different spatial scales:

Sioux Falls, South Dakota
Richmond, Virginia
Chesapeake Bay Area
Utah
Western Oregon

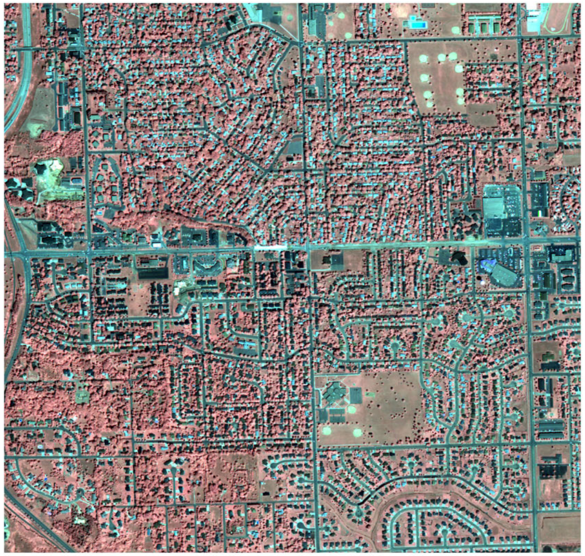
Data:

- training/testing data:

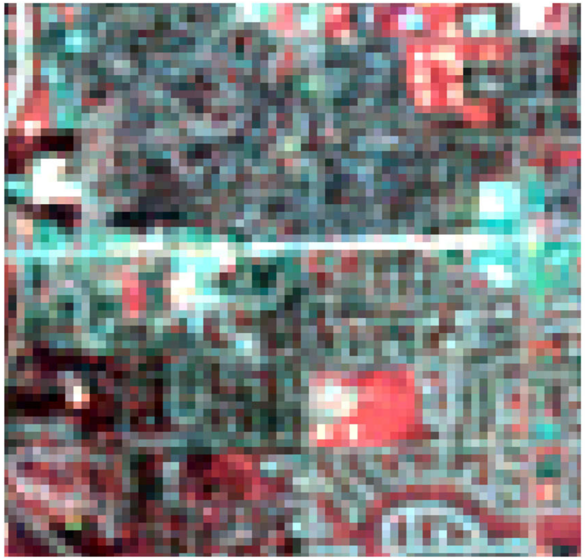
  IKONOS, NASA SDP
  Digital Orthophoto Quadrangles (DOQ), USGS

- mapping:
  Landsat 7 ETM+

IKONOS

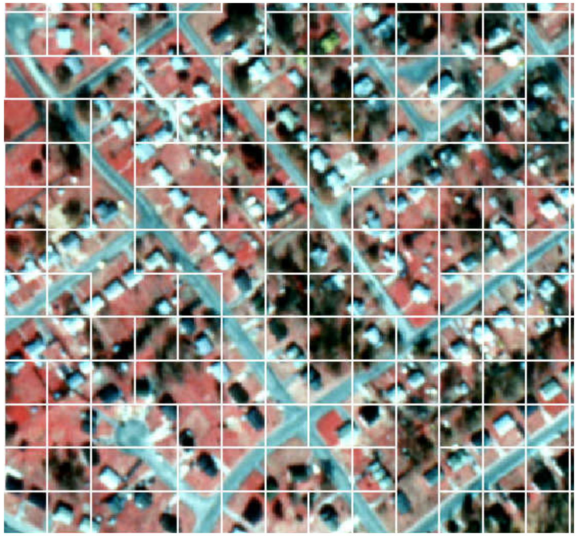


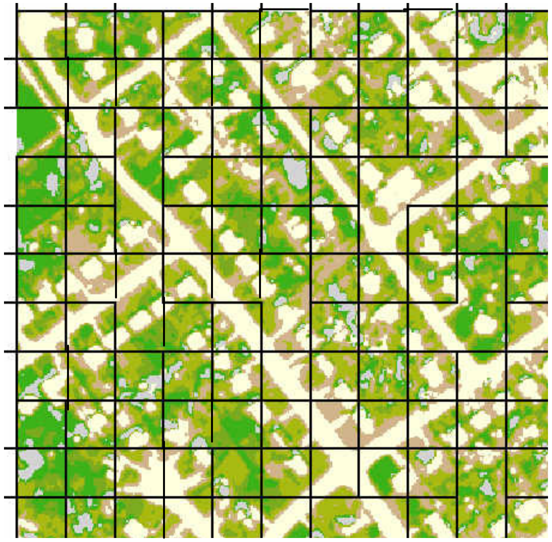
Landsat 7 ETM+



Color Composite Imagery of SE Sioux Falls, SD

Step 1. Estimate of % impervious surface from high resolution data (e.g. IKONOS)

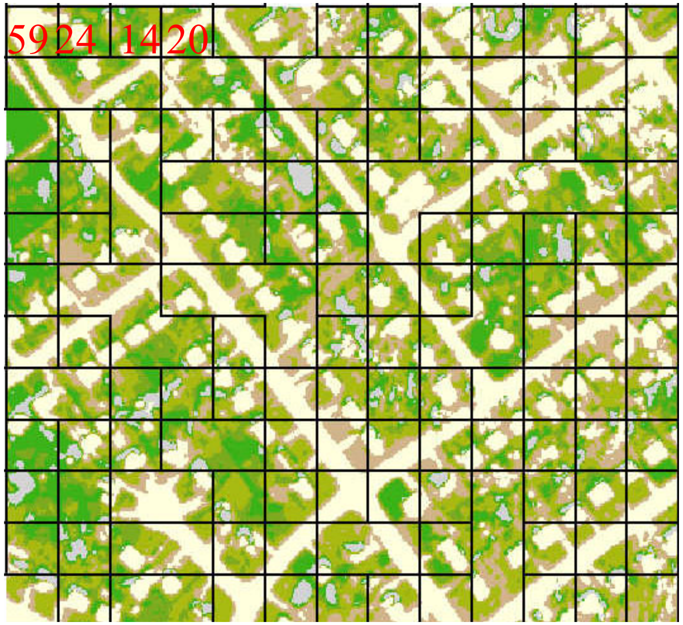


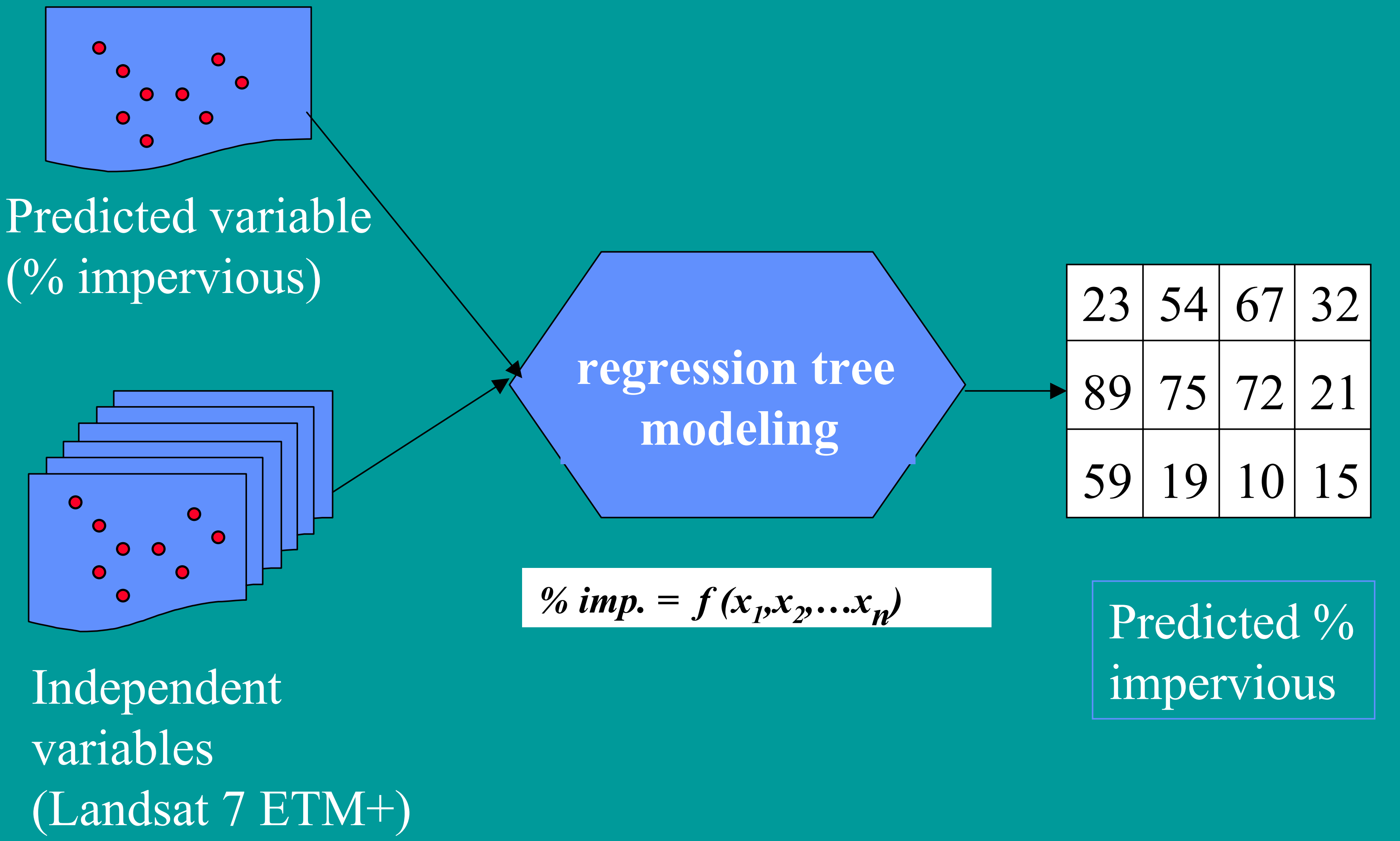


Step 2. Establish a training sample dataset using Landsat 7 ETM+ and estimated % impervious surface

TM bands and transformation (30m)

Estimated % impervious surface from IKONOS




Predicted variable
(% impervious)
Independent
variables
(Landsat 7 ETM+)
regression tree
modeling
% imp. = f (x1,x2,…xn)
23 54 67 32
89 75 72 21
59 19 10 15
Predicted %
impervious

Model

Rule 1: [12 cases, mean 20.4366207, range 0.288889 to 49.55042, est err 10.9970322]

if

tmband4 > 61

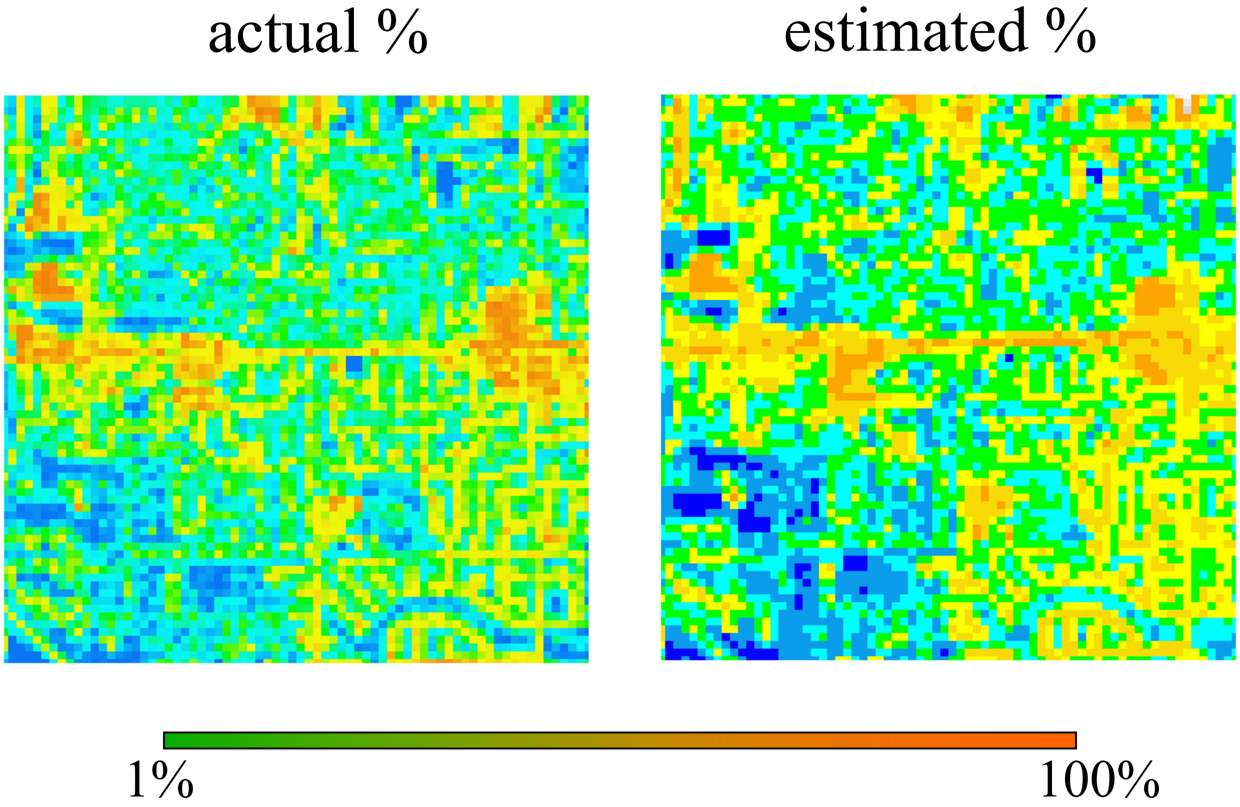
NDVI > 0.0619469

then

percent impervious = 88.3936 - 1.016 tmband4 + 0.44 tmband3 - 31.7 NDVI

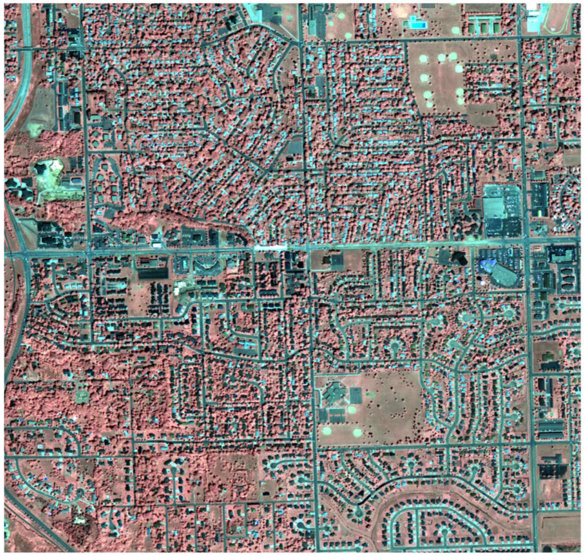
Rule 2: …,…

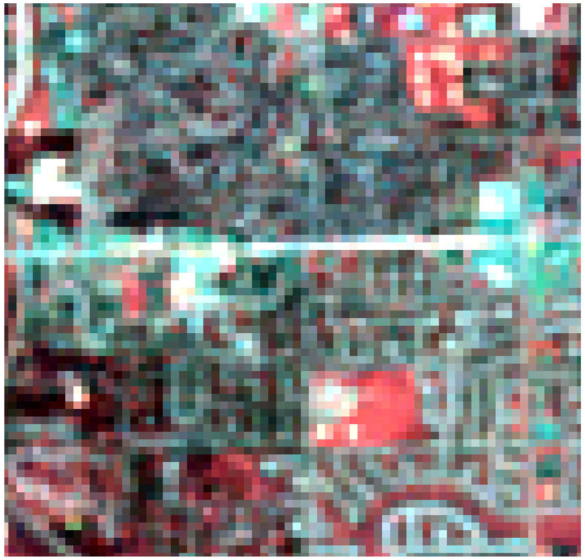
Impervious Surface of SE Sioux Falls, SD



IKONOS

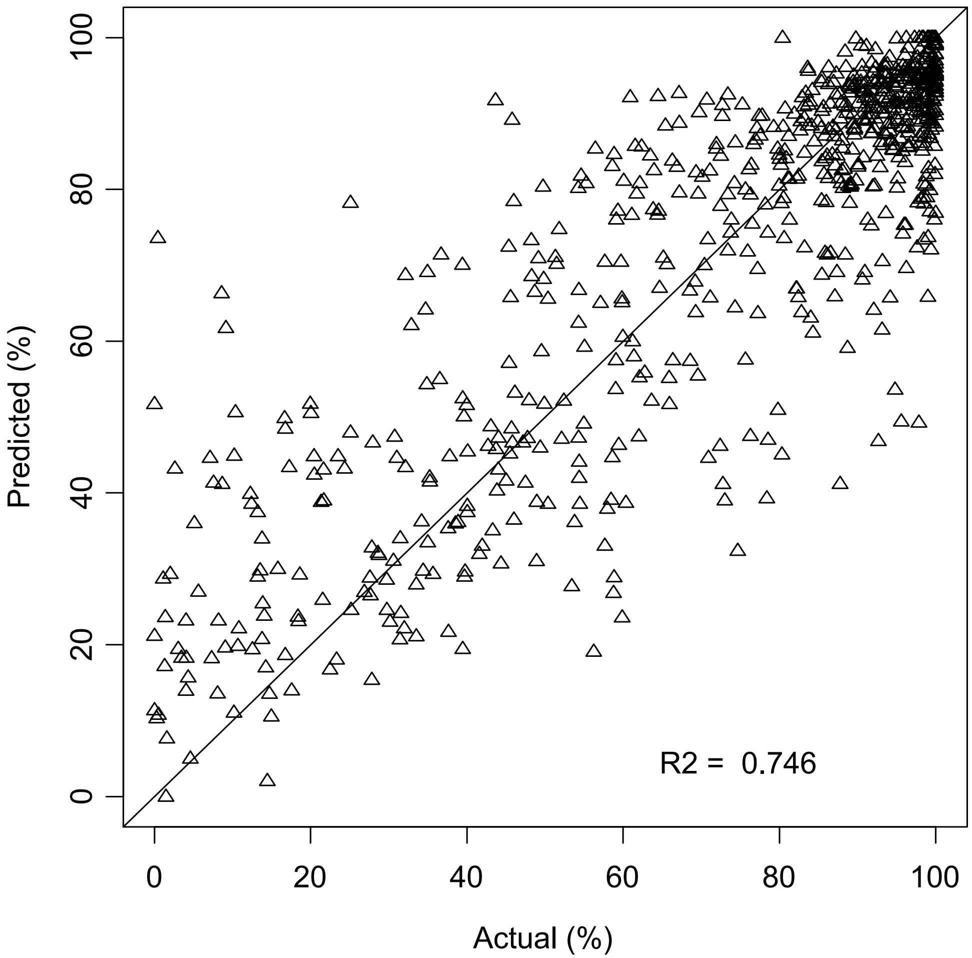
Landsat 7 ETM+

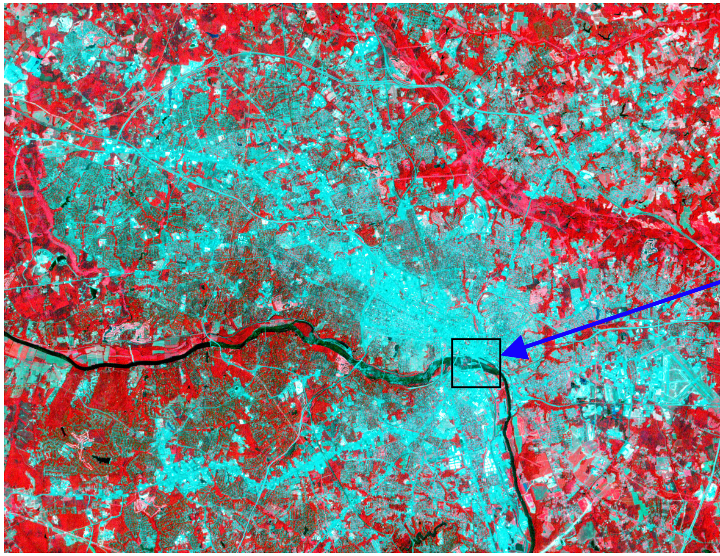




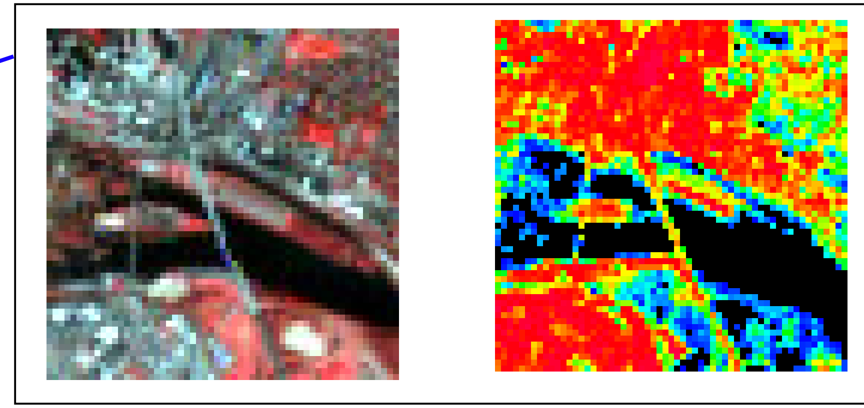
Color Composite Imagery of SE Sioux Falls, SD

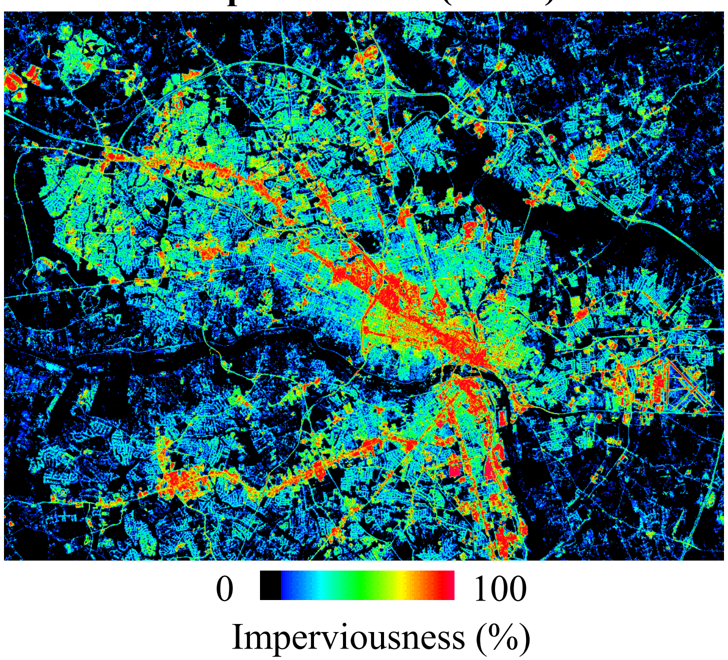
Land Cover Characterization Program





Richmond, VA, ETM+ image (above) and estimated imperviousness (below)



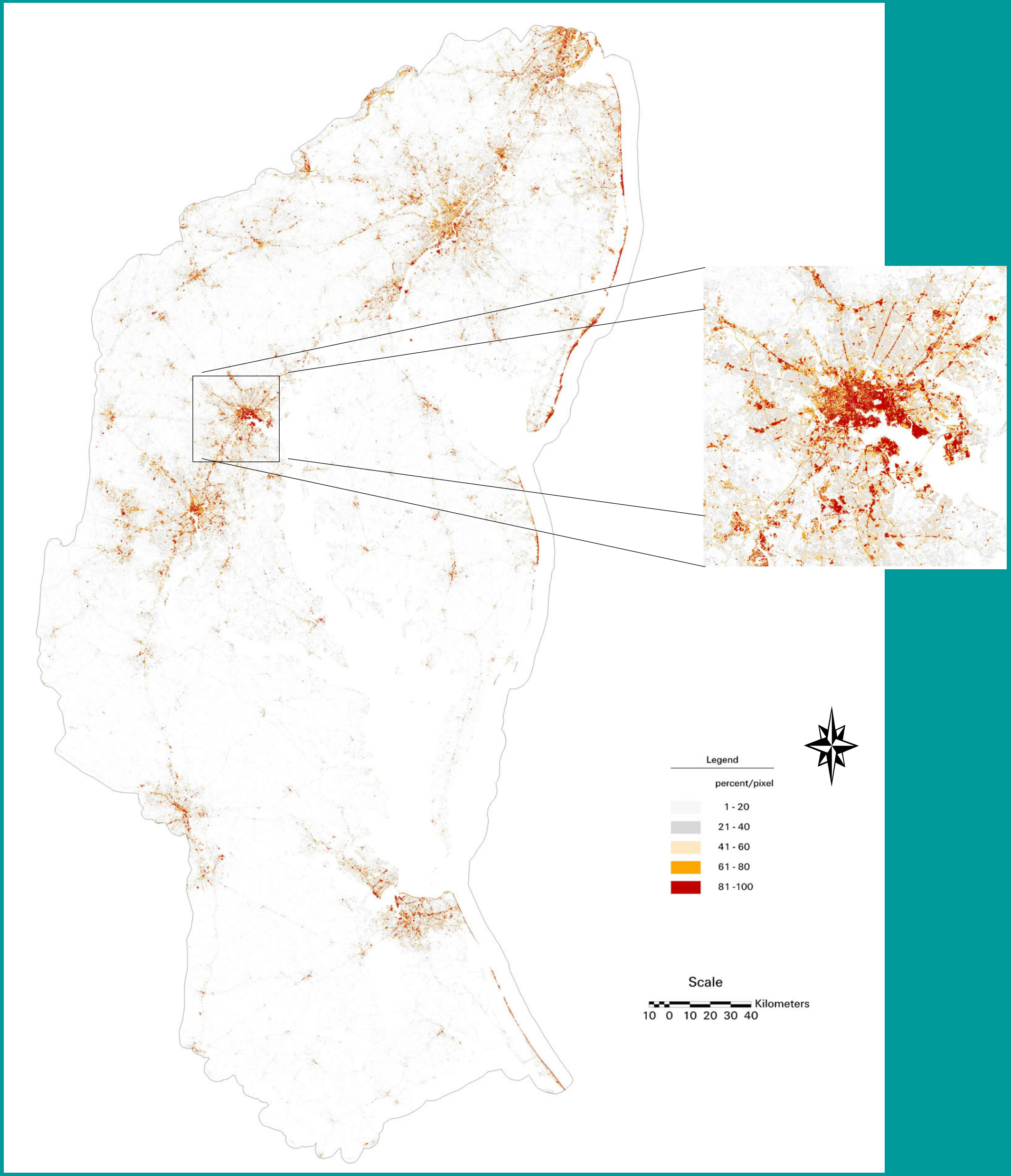


Estimating imperviousness from Landsat 7 ETM+ image using a regression tree method

USGS

Land Cover Characterization Program

Chesapeake Bay Study Area



Results:

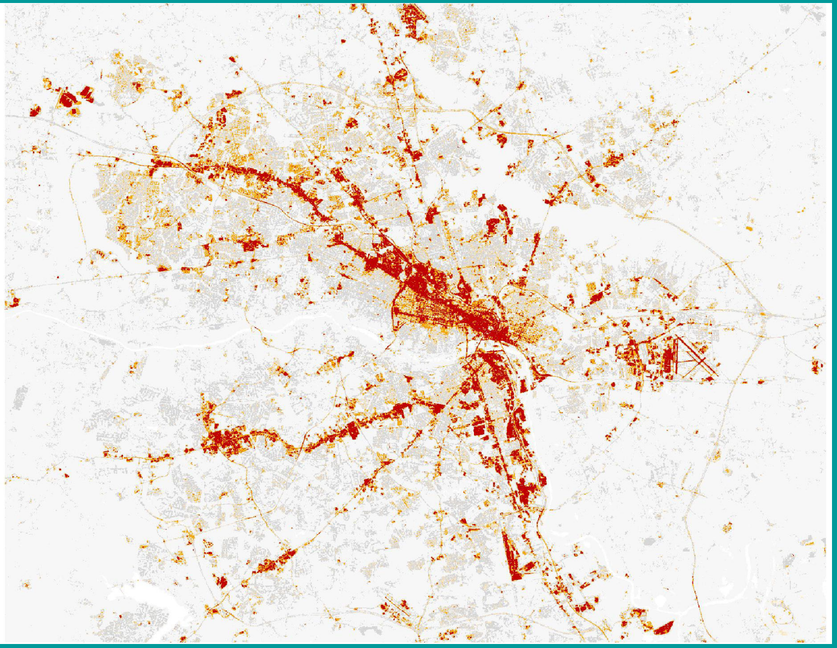
mostly selected variables (bands)

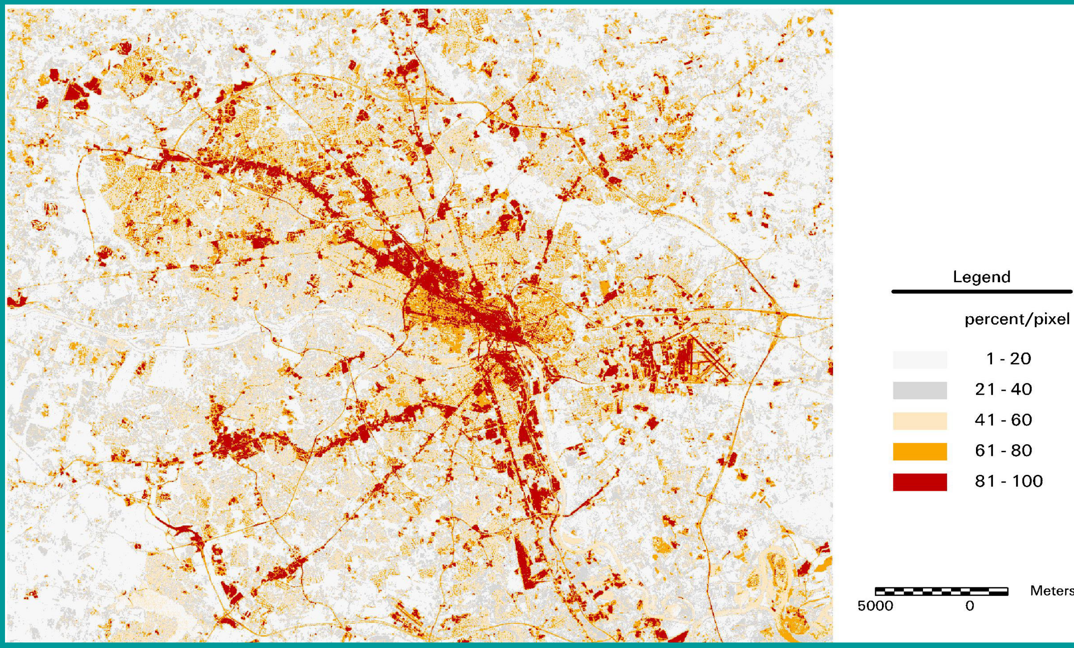
- Tasseled-cap transformation plus thermal
  or
  band 4 (NIR), band 7 (mid-IR) and band 3 (VIS) plus thermal

- leaf-on or leaf-on and leaf-off imagery

Location	MAE(%)	*r*	Variables
Sioux Falls, SD	**9.6**	**0.88**	Leaf-on greenness, band 3, 4, 7 and thermal
Richmond, VA	**9.1**	**0.90**	leaf-on 1,4,5,7 and thermal
Chesapeake Bay area	**9.3**	**0.88**	leaf-on and leaf-off Tasseled-cap transformation bands and thermal

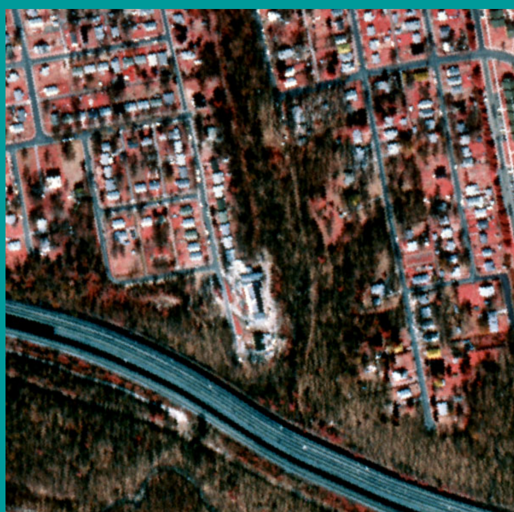
Comparison of using two prediction models



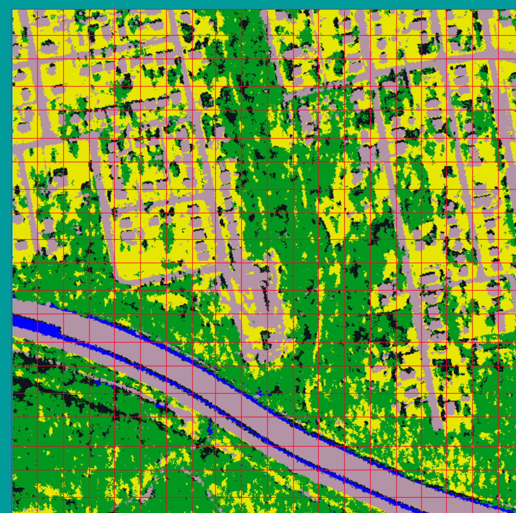


Mapping sub-pixel percent tree canopy density

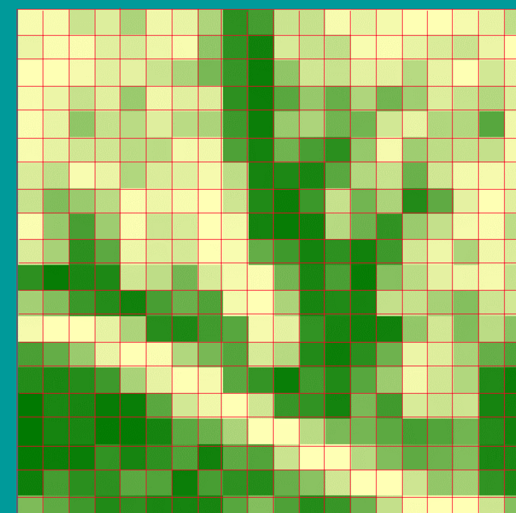
Reference data development
– from color infrared DOQ / IKONOS



DOQ image (1m)



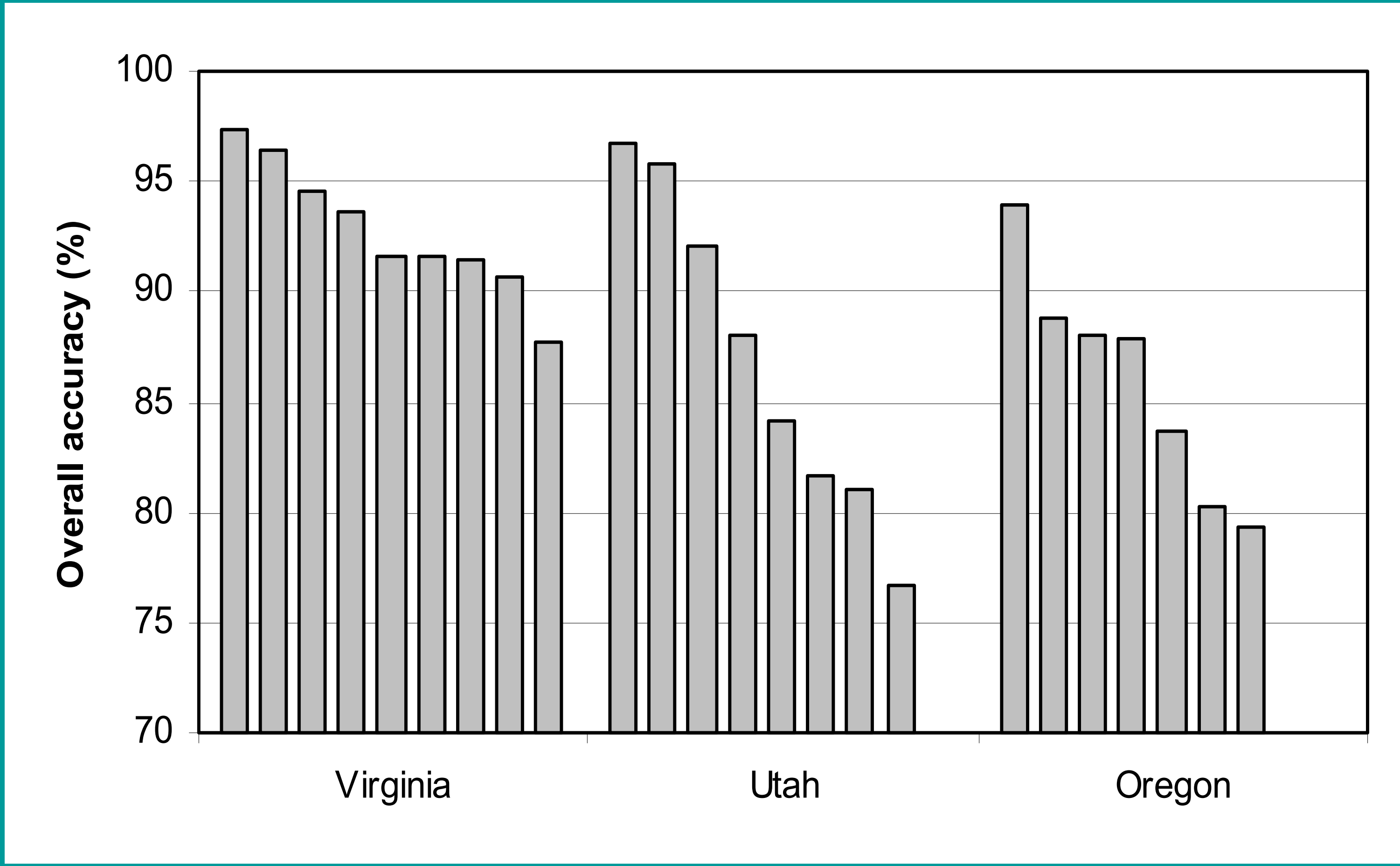
Classification (1m)



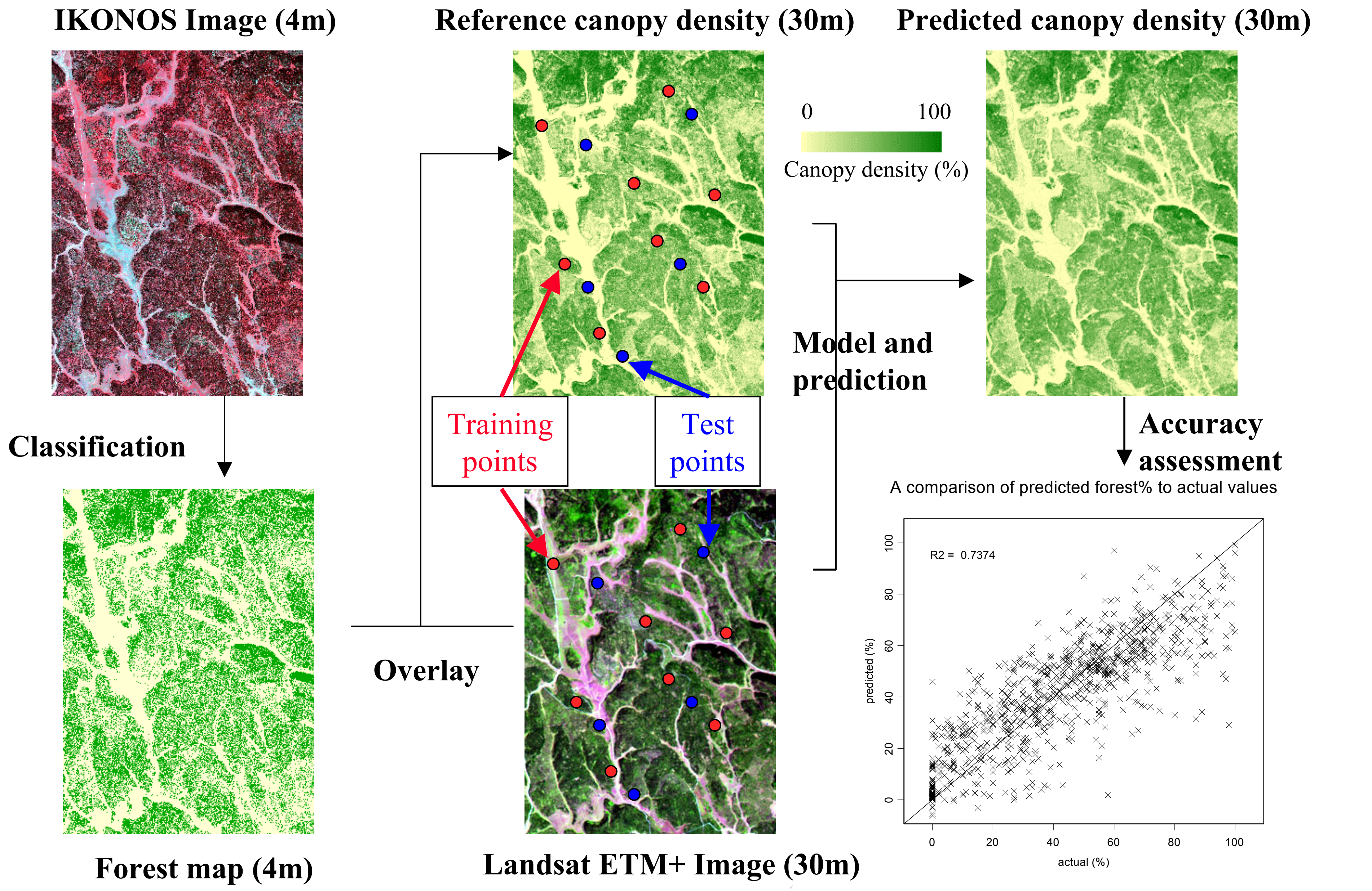
Canopy density (30m)

0 100%

Figure 2. Five-fold cross validation estimates of the accuracy for the decision tree classification of DOQ images. Each bar represents the estimated accuracy of classifying one DOQ image window.



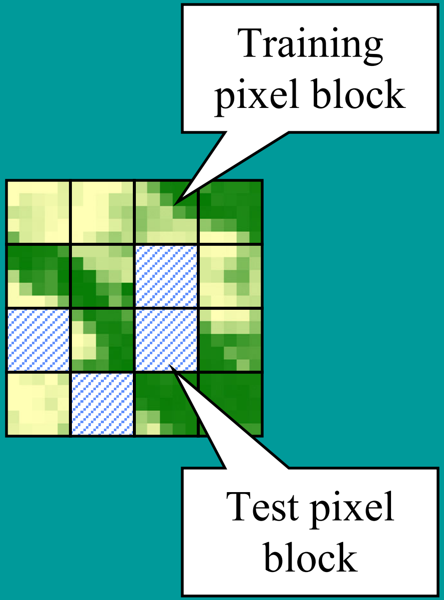
Land Cover Characterization Program



Model development

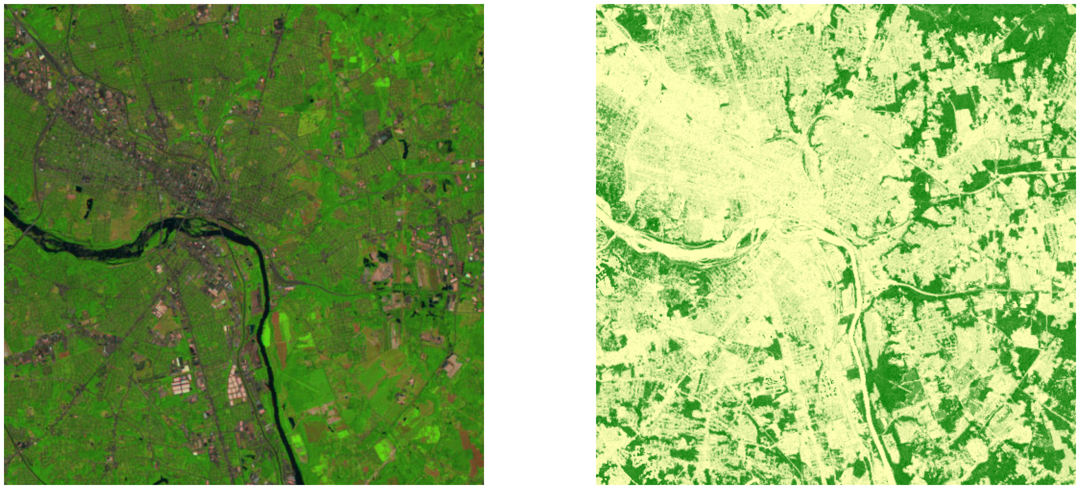
– Splitting reference data for training and accuracy assessment

- Pixel-based random sampling
  - Strong spatial auto-correlations between training and test samples
  - Accuracy estimates inflated
- Block-based random sampling
  - Reference image divided into equal-sized blocks
  - Randomly select some blocks for training/test
  - Reduce spatial auto-correlation
  - Accuracy estimates more realistic

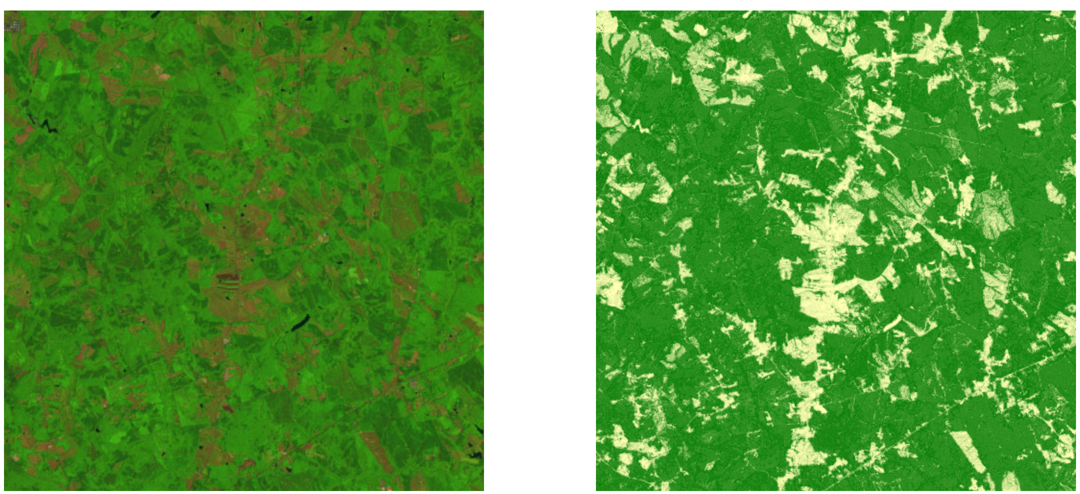


Estimated tree canopy density in two areas of Virginia

Landsat 7 ETM+ image | Estimated canopy density



(a) Richmond



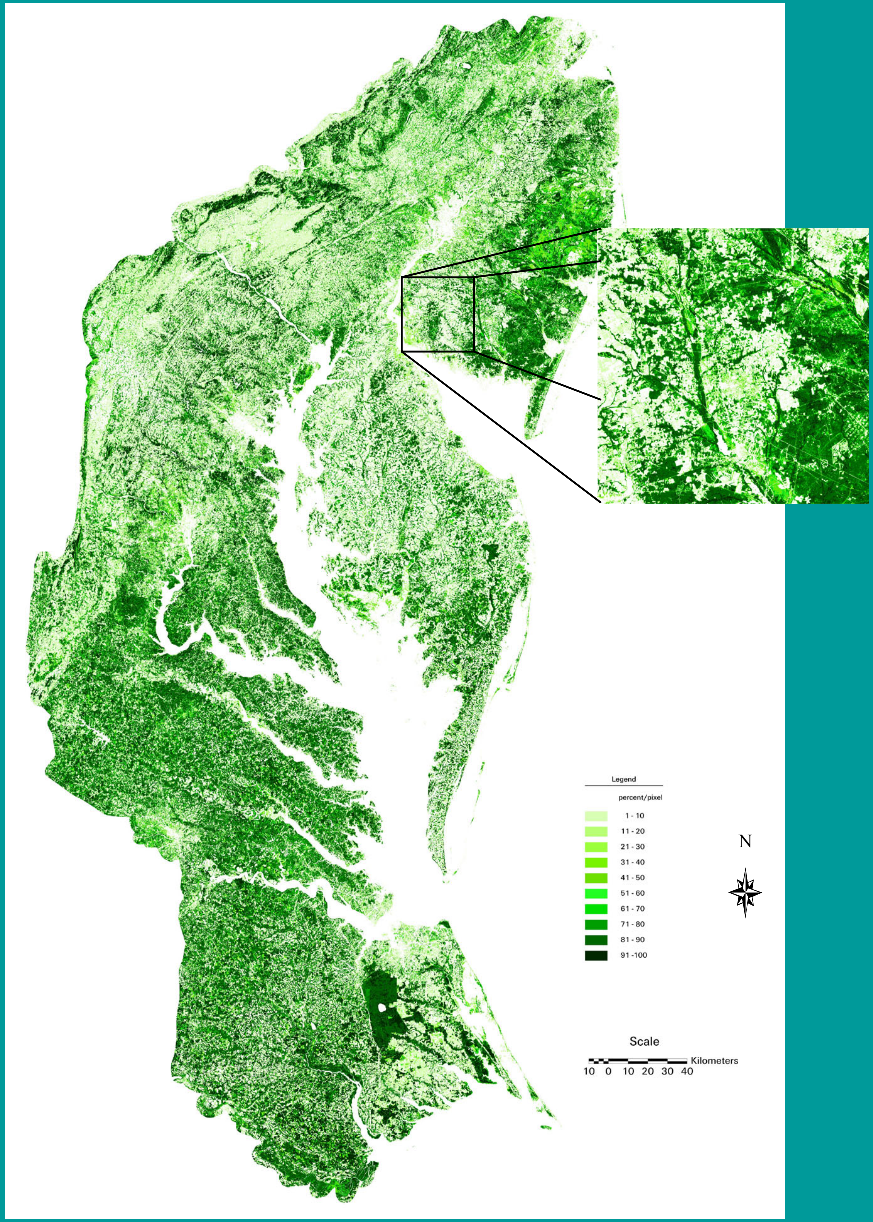
(b) Cumberland State Forest

tree canopy density (%)

Table 2. Mean absolute difference (*MAD*) and correlation (*r*) between predicted and actual canopy density values on independent test samples. The unit of *MAD* is tree canopy density in percentage.

Study area (two-scene mosaic)	Regression tree model		Linear regression model	
	MAD (%)	*r*	*MAD* (%)	*r*
Virginia	11.65	0.89	13.15	0.83
Utah	9.92	0.85	10.14	0.70
Oregon	10.98	0.87	11.93	0.80

Chesapeake Bay Study Area



U. S. Geological Survey
EROS Data Center
Sioux Falls, SD 57198

Conclusions (impervious surface)

- for three area tested, the regression tree was capable of predicting imperviousness with consistent and acceptable accuracy (MAE ~ 10% and r ~ 0.9)
- the most relevant set of input variables in model prediction were one band each in visible, NIR,mid-IR and thermal-IR or the three Tasseled-cap bands
- spatial extensibility of predictive model can be beneficial in large-area impervious surface mapping

Conclusions (tree canopy density)

- for three area tested, the regression tree prediction was reasonable (MAE ~ 11% and r ~ 0.85)

- the independent variables were Landsat 7 ETM+ seven bands of two images (leaf-on and leaf-off)

Conclusions (cont.)

- For large-area impervious surface mapping, collecting field-based measurements for training/test data is cost-prohibitive. High-resolution imagery provides an alternative.

- the validation data should be independent from the training data to reduce spatial auto-correlation

Factors effecting model prediction:

- image co-registration
- interpretability of high resolution data
- temporal consistency of data sources
- spectral confusion

Future work

- Uncertainties in reference data
  - Temporal difference between high res. Images and ETM+ images
  - Misclassification error
- Feature selection
  - Use most relevant variables
  - Develop more compact model
- Non-forest mask
  - Reduce commission error over non-forest areas

USGS EROS Data Center (EDC)

http://edc.usgs.gov/